DEFLOG: on the logical interpretation of prima facie justified assumptions

Bart Verheij

Department of Metajuridica, Universiteit Maastricht
P.O. Box 616, 6200 MD Maastricht, The Netherlands
bart.verheij@metajur.unimaas.nl, http://www.metajur.unimaas.nl/~bart/

Abstract

Assumptions are often not considered to be definitely true, but only as prima facie justified. When an assumption is prima facie justified, there can for instance be a reason against it, by which the assumption is not actually justified. The assumption is then said to be defeated. This requires a revision of the standard conception of logical interpretation of sets of assumptions in terms of their models. Whereas in the models of a set of assumptions, all assumptions are taken to be true, an interpretation of prima facie justified assumptions must distinguish between the assumptions that are actually justified in the interpretation and those that are defeated.

In the present paper, the logical interpretation of prima facie justified assumptions is investigated. The central notion is that of a *dialectical interpretation* of a set of assumptions. The basic idea is that a prima facie justified assumption is not actually justified, but defeated when its so-called *dialectical negation* is justified. The properties of dialectical interpretation are analyzed by considering partial dialectical interpretations, or *stages*, and by establishing the notion of *dialectical justification*. The latter leads to a characterization of the existence and multiplicity of the dialectical interpretations of a set of assumptions. Since dialectical interpretations are a variant of stable semantics, the results are relevant for existing work on nonmonotonic logic and defeasible reasoning, on which the present work builds.

Instead of focusing on defeasible rules or arguments, the present approach is sentence-based. A particular innovation is the use of a conditional that is prima facie justified (just like other assumptions) instead of an inconclusive conditional.

1 Introduction

When someone is arrested, he is assumed to be innocent.[1] When an object looks red, it is assumed that it therefore is red.[2] Both assumptions are not considered to be definitely true, but are only taken as prima facie justified. Additional information can have the effect that such prima facie justified assumptions are actually not justified. Someone is not actually held innocent, when his guilt is proven by law. When the object is illuminated by a red light, it is not actually taken for granted that it is red since it looks red.

Such prima facie justified assumptions that are not always actually justified, are the topic of investigation of the present paper. The research reported on here builds on and extends previous work on nonmonotonic logic and defeasible reasoning and argumentation. The topic of prima facie justified assumptions has not received much attention since previous work typically focuses on defeasible rules or arguments and not on defeasible statements.

The work of Reiter, Hage, Prakken, Pollock, Vreeswijk, Loui, Dung and Toulmin[3] is among the most influential for my thinking about the subject. A lot of other research (in logic, artificial intelligence, argumentation theory and law) is relevant and has influenced me in ways that are less obvious. There are good overviews of such research (e.g., Haack 1978, Ginsberg 1987, Gabbay, Hogger & Robinson 1994, Read 1995, Van Eemeren, Grootendorst & Snoeck Henkemans 1996, Bench-Capon 1997, Hage 2000, Chesñevar, Maguitman & Loui 2000, Prakken & Vreeswijk 2002).

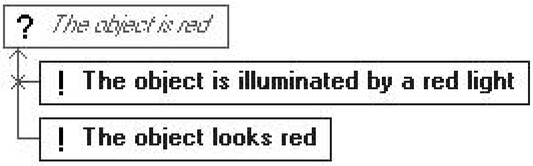
[1] Cf. a basic principle of criminal procedure: the *presumptio innocentiae* (codified in, e.g., article 11 of the Universal Declaration of Human Rights). The example is also used by Bondarenko, Dung, Kowalski & Toni (1997).

[2] Cf. Pollock's basic example of an *undercutting defeater* (e.g., Pollock 1987, 1995).

[3] The order in which the names appear only reflects the accidental chronology of my intellectual history. Some relevant sources are Reiter's (1980), Hage's (1996, 1997), Prakken's (1997), Pollock's (1995), Vreeswijk's (1997), Loui's (1998), Dung's (1995) and Toulmin's (1958).

1.1 Personal motivation

The interpretation of prima facie justified assumptions is analyzed in terms of the logical system DefLog, that is presented here.[4] The development of this logic was amongst others guided by my research on the automated argument assistant ArguMed (Verheij 1999), a prototypical computer program that assists the user while doing argumentation.[5] The 1999 version of ArguMed (ArguMed 2.0) focused on undercutting defeaters (Pollock 1987, 1995), i.e., reasons that block the connection between a reason and its conclusion. For instance, in ArguMed, the red object example looks as follows:



? *The object is red*

! **The object is illuminated by a red light**

! **The object looks red**

The exclamation marks indicate assumptions, the question mark an issue. While the assumptions that the object looks red and that the object is illuminated by a red light are justified (indicated by the dark bold font), the conclusion that the object is red is not justified (indicated by the light italic font) since the prima facie reason that the object looks red is not justifying.

I had three main reasons for the further development of the argumentation theory of ArguMed 2.0. The initial reason was that its expressiveness had some obvious limitations. In fact, a generalization of the argumentation theory underlying ArguMed 2.0 in the end led to DefLog.[6] Especially, ArguMed's argumentation theory only allowed the attack of the connection between a reason and its conclusion (by its focus on undercutting defeaters), and not the attack of ordinary statements. Moreover, it treated the support of the connection between a reason and its conclusion (cf. Toulmin's (1958) warrants and backings) different from the support of ordinary statements. The distinction seemed to be an artifact of the logical language used in the argumentation theory. As a consequence, I started using a simple logical language with two conditionals (denoted as ⤳ and ⋊ respectively), one expressing the support of a conclusion by a reason, the other expressing the attack of a conclusion by a reason. The resulting language turned out to be surprisingly expressive considering its simple structure, mainly because the conditionals are treated as object-level connectives that can be nested. The result was that reasons for and against a conditional relation (cf. Toulmin's warrants and Pollock's undercutting defeaters, respectively) became special cases of reasons for and against statements in general.

The second main reason was that I wanted to investigate to what extent the dialectical arguments in the style of ArguMed (essentially trees that consist of statements with reasons for and against them, and in which the connecting arrows can themselves have reasons for and against them) could count as a generalization for defeasible reasoning of the proofs of deductive reasoning. The guiding idea was that there should be a close relation between being the justified conclusion of an evaluated dialectical argument and being actually justified, i.e., being justified in a dialectical interpretation. The investigation led me to adapt my naive notion of dialectical arguments (see Verheij 2000a, sections 2 an 10) and to the discovery of the notion of dialectical justification. The latter notion resulted in an elegant characterization of the existence and multiplicity of the dialectical interpretations of sets of prima facie justified assumptions. It turned out that a slight generalization of the logical language in order to express a statement's defeat (by using the connective × for dialectical negation) considerably simplified the definitions and resulted in a clear view on the interpretation of prima facie justified assumptions. The combination of × and ⤳ allowed the elimination of ⋊: it could be replaced by ⤳×.

The third reason was that I wanted to get a good view on the nature of logic in the context of defeasible reasoning. Many of the existing theories obviously described some of the 'right' concepts relevant for defeasible reasoning - I have already cited some of the best sources -, but not in a way that in my opinion sufficiently clarified the relation with standard logical concepts, such as models of a theory, valid consequence and proof. What I found especially confusing was the fact that several approaches used

[4] For earlier publications on DefLog, see Verheij (2000b, 2001b, 2002). All draw on the manuscript by Verheij (2000a).

[5] See also http://www.metajur.unimaas.nl/~bart/aaa/.

[6] After ArguMed 2.0, a version of ArguMed has been developed that is based on DefLog. See the paper by Verheij (*to appear*) on argument assistants.

separate logical layers, one for the basic information, another for the information that led to defeasibility. The normal separation between a logical object language and the meta-language in which the logic is described was regularly extended with an intermediate language for the defeasibility information. In particular, defeasible conditionals were often expressed outside the logical object language,[7] as were the priority or defeat information.[8] However, to obtain the required effect, the different layers of course had to be somehow connected.[9] I have always found that this resulted in confusing formalizations, that obscured the relation between 'deductive' and 'defeasible' logic.[10] The separation seemed to be dictated by the fact that defeat was treated as a property of arguments (roughly in the sense of derivations). Instead DEFLOG is sentence-based and focuses on the defeat of prima facie justified assumptions, instead of on the defeat of the arguments in terms of them.

DEFLOG's basic concepts have been chosen to be as close as possible to basic concepts of standard logic. With some exaggeration, the development of DEFLOG has led me to believe that the basic difference between a deductive and a defeasible logic is dialectical negation paired with dialectical interpretation.

1.2 The formally related work of Dung, Bondarenko, Kowalski and Toni

DEFLOG is part of a long history. Its concepts and formal techniques have been especially influenced by the abundance of work on logics modeling defeasible argumentation (see for instance the overview by Prakken & Vreeswijk 2002).

Particularly relevant is the work by Dung (1995) on argumentation frameworks and by Bondarenko, Dung, Kowalski & Toni (1997) on assumption-based frameworks. Conceptually there is an important difference in starting point, since argumentation frameworks and assumption-based frameworks focus on arguments, while DEFLOG is sentence-based. However, formally there are close relations. DEFLOG's central definition - viz. that of a dialectical interpretation of sets of prima facie justified assumptions - corresponds to their stable semantics, which is closely related to the stable models of logic programming (Gelfond & Lifschitz 1988). An important formal difference between DEFLOG and Dung's (1995) argumentation frameworks is the richer language used by DEFLOG. Essentially, Dung uses a language that can only express a fixed attack relation, whereas in DEFLOG's language the attack relation is not fixed, but can depend on the other information, and is moreover flanked by a support relation.

The expressiveness of DEFLOG's language is also a difference with Bondarenko, Dung, Kowalski & Toni's (1997) assumption-based frameworks. The latter use a fixed set of rules of inference, whereas DEFLOG uses object-level conditionals that can depend on (i.e., be derived from) the other information. The contraries that occur in assumption-based frameworks (and that should not be confused with their non-provability claims) play a role that is related to DEFLOG's dialectical negations. However, the contraries are not expressed using an object-level connective (as DEFLOG's dialectical negations), but by a meta-level mapping of the language into itself. Formally, such a mapping can do the job, but I believe that DEFLOG's use of a dedicated connective tends to illuminate what is going on.[11] Bondarenko, Dung, Kowalski & Toni's (1997) application of contraries is very different from that of DEFLOG's dialectical negations. Whereas contraries are used as a technical tool for the reconstruction of related logical

[7] E.g., Reiter's (1980) defaults, Pollock's (1995) prima facie reasons, Nute's (1994) defeasible rules, Vreeswijk's (1997) defeasible rules of inference, Prakken's (1997) defeasible rules, Bondarenko, Dung, Kowalski & Toni's (1997) rules of inference, but not Hage's (1997) rules.

[8] E.g., Pollock's (1995) defeaters, Nute's (1994) defeaters, Vreeswijk's (1997) conclusive force relation, Prakken's (1997) priority relation, Bondarenko, Dung, Kowalski & Toni's (1997) contrary mapping.

[9] I know of only one logical formalism, viz. Hage's (1996, 1997) Reason-Based Logic, to which I have contributed (Verheij 1996b), that uses an integrated language. However, the formalization of Reason-Based Logic is so different from standard logics (for instance by its typical predicate and function symbols and its denotation of facts as terms) that it is hard to see the connection.

[10] A logic is here any formalization of concepts related to reasoning. When I speak of standard logics, I think in the first place of classical propositional and first-order predicate logic. A deductive logic is a logic that is based on truth-preservation or on deduction rules (that are not defeasible). A defeasible logic is a logic in which assumptions, rules of inference or derivations are somehow defeasible.

[11] During the development of DEFLOG I was well aware of the close formal connection with Dung's (1995) work. The formal connection with the work of Bondarenko, Dung, Kowalski & Toni (1997) only dawned on me when DEFLOG was already finished.

formalisms that mainly relies on a second tool, viz. non-provability, dialectical negations are used to express the defeat of a prima facie justified assumption.

Next to the differences in conceptualization, expressiveness and application, in the present paper, techniques are used that differ from those used by Dung, Bondarenko, Kowalski and Toni. Especially, the investigation of the stages of a set of assumptions (essentially the dialectical interpretations of subsets of the assumptions) and the definition of dialectical justification (a variant of the notion of admissibility that is at the heart of the work of Dung, Bondarenko, Kowalski and Toni, but with nicer properties) are new. These techniques are directly applicable to the argumentation and assumption-based frameworks of Dung, Bondarenko, Kowalski and Toni (see also section 6).

A final difference is that, as said, DEFLOG has been designed in an attempt to stay as close as possible to concepts of standard logics, thereby hopefully illuminating the relation between deductive and defeasible logic.

1.3 Some key notions of DEFLOG

DEFLOG's logical language has two connectives $\times$ and $\rightsquigarrow$. The former denotes *dialectical negation*, the latter *primitive implication*. Dialectical negation expresses defeat. The dialectical negation $\times\varphi$ of a sentence φ expresses that the statement that φ is defeated. Dialectical negation is the basic logical tool to deal with the interpretation of prima facie justified assumptions. When the dialectical negation of a prima facie justified assumption is (actually) justified, the assumption is not actually justified, but defeated. The properties of dialectical negation are significantly different from standard negation (as will be discussed below). Note that dialectical negation is not meant to replace standard negation, but introduced as a different concept. As a result, it makes sense to use standard and dialectical negation side-by-side in one language. (Cf. the discussion of the Nixon diamond in section 2 below.)

Primitive implication is intended to express elementary conditional relations as they exist contingently in the world. Examples of primitive implication are 'If an object looks red, it is red' and 'If John is a thief, he is punishable'. This is in contrast with the material implication of classical logic. Of course the material implication can be used to express elementary, contingent conditional relations, but the material implication is also intended to express tautologous conditional relations (cf. its use in the deduction theorem and the well-known paradoxes of the material implication) and 'redundant' conditional relations such as 'If John is a thief and it rains, John is punishable'. (See, e.g., Haack 1987 and Read 1995 for a discussion of the material implication and what it represents.) The properties of the primitive implication built into DEFLOG are sharply delimited: it only validates Modus ponens (From $\varphi \rightsquigarrow \psi$ and φ, conclude ψ). Primitive implication does for instance not in general validate the classical introduction rule for conditionals (Given a proof of ψ assuming φ, obtain a proof of $\varphi \rightsquigarrow \psi$ that does not assume φ). Notwithstanding the delimitation of its properties, the use of primitive implication gives DEFLOG adequate expressiveness.[12] Differences between primitive and material implication are discussed below.

The central definition of DEFLOG is that of the *dialectical interpretation* of a set of prima facie justified assumptions. There are two main differences between the standard logical interpretation of sets of assumptions in terms of their models and the interpretation of sets of assumptions in terms of their dialectical interpretations. The first is that, in the standard models of a set of assumptions, all sentences are taken to be true: all assumptions are assigned the same positive status, in logic usually referred to as 1 or t. This corresponds to the idea of taking the assumptions as definitely true. A model of a set of assumptions is then a logically possible world in which all assumptions are true. In the dialectical interpretation of a set of assumptions in DEFLOG, however, not all sentences need to be given a positive evaluation: an assumption can be either positively evaluated, viz. as *justified*, or negatively, viz. as *defeated*, formally referred to as j and d, respectively. This corresponds to the idea of taking the assumptions as prima facie justified, instead of definitely true: some of the prima facie justified assumptions turn out to be actually justified, others as defeated in the dialectical interpretation. The actually justified assumptions defeat the other assumptions. Which assumptions are actually justified and which defeated is essentially constrained as follows: in a dialectical interpretation, an assumption is defeated if and only if the assumption's dialectical negation follows from the assumptions that are justified.

[12] For instance, every logic that has a Hilbert-style proof theory, i.e., one that uses axioms and one rule of inference, viz. Modus ponens, can in a trivial way be mimicked using primitive implication. Many logics have such a proof theory.

The second main difference between the models of standard logic and DefLog's dialectical interpretations is that in the models of standard logic, the whole language is interpreted, i.e., all sentences of the language are assigned a status (usually either true or false), while in dialectical interpretations, this need not be so: a dialectical interpretation has an extent, that consists of the sentences of the language that are assigned a status. The intuitive idea is that in a dialectical interpretation only those sentences are evaluated as are justified or defeated by the theory. In fact, the analogy between dialectical interpretations and sets of consequences of consistent sets of sentences is closer than that between dialectical interpretations and models. This explains the 'partiality' of dialectical interpretations.

1.4 Informal examples of DefLog

Before discussing further details of DefLog, let's consider the two examples of the start of the introduction (section 1): the presumption of innocence and the red-looking object. How are they to be logically analyzed?

Example (1.1): the presumption of innocence

There are two prima facie assumptions:

```
innocent
proven_guilty → ×innocent
```

The first sentence expresses the assumption of innocence, the second that when guilt is proven (by proof in the legal sense, not in the logical sense), the assumption of innocence is defeated. Given these two assumptions, the assumption of innocence is not only prima facie justified, but also actually. There is no information that can lead to the defeat of one of the prima facie justified assumptions: the antecedent of the conditional is not satisfied. When however a third assumption `proven_guilty` is added, it follows that `×innocent`. When the assumptions would be taken as definitely true, an inconsistency arises: both `innocent` and `×innocent` follow. Since the assumptions are interpreted as being prima facie justified, the situation is different. The prima facie assumption of innocence is countered by its dialectical negation. As a result, the prima facie assumption `innocent` is not actually justified, but defeated. Note that dialectical negation is inherently 'directed', in the following sense. Since `×innocent` follows, `innocent` is defeated. However, it is not the case that since `innocent` is prima facie justified, `×innocent` is defeated. This is in contrast with standard negation where the truth of a negated sentence, implies the sentence's falsity, while also a sentence's truth implies the falsity of its negation.[13]

Example (1.2): the red-looking object (Pollock 1987, 1995)

Initially, there are two prima facie assumptions:

```
looks_red
looks_red → is_red
```

The former expresses that some object looks red, the latter that if an object looks red, it is red. It follows that the object looks red. However, let's make two additional prima facie assumptions:

```
red_light
red_light → ×(looks_red → is_red)
```

The first sentence expresses that the object is illuminated by a red light. The second expresses that if an object is illuminated by a red light, it is defeated that the object is red when it looks red. When the four sentences are together assumed to be prima facie justified, the prima facie assumption that the object is red when it looks red, is defeated, and it does not follow that the object looks red. Again, an inconsistency would arise when the assumptions would not be taken as prima facie justified, but as definitely true.

[13] Bondarenko, Dung, Kowalski & Toni (1997) also analyze the presumption of innocence. However, they do not analyze the presumption of innocence *per se*, but instead the conditional expression that someone is innocent unless proven guilty. To this effect they use a weak negation expressing (logical, not legal) non-provability, and a generic non-provability assumption that can be defeated. When ~ denotes non-provability, they use a conditional of the form `~guilty → innocent` to formalize the example. Since `~guilty` is defeasibly assumed (just like every other statement of the form ~φ), the innocence follows defeasibly. Note that in this way the presumption of innocence is a conditional with a defeasibly fulfilled antecedent and not a separate, prima facie justified assumption, as in the analysis here.

Both examples show DefLog’s nonmonotonicity: after adding assumptions, an initial consequence no longer follows.

1.5 The contribution of the present research

Among the innovations of DefLog and the contributions of this paper are the following:

- The investigation of the logical interpretation of *prima facie justified assumptions* (in terms of dialectical interpretations) in contrast with the standard logical interpretation of definitely true assumptions (in terms of models).
- The design of a *sentence-based* theory of defeasible reasoning instead of a rule-based or argument-based theory.
- The definition and analysis of *dialectical negation* and *primitive implication*, and the discovery that attack and several other notions from defeasible logic (like Toulmin’s warrants and Pollock’s undercutters) can be analyzed in terms of dialectical negation and primitive implication.
- The distinction of *two kinds of defeasibility for conditionals*: being prima facie justified and being inconclusive. Whereas the former type applies to any assumption, the latter is restricted to conditionals. DefLog is based on a prima facie justified conditional.
- The discovery of the notion of *dialectical justification* and its relation to the *existence and multiplicity problems* for dialectical interpretations, and its subtle distinction from the notion of admissibility.
- The use of *genuine sentential connectives* ×, ⇝ and ⋊, allowing nested expressions, in the context of defeasible argumentation, in an attempt to normalize the expressiveness of logics for defeasible argumentation.
- The notion of *stages as partial interpretations* of sets of prima facie justified assumptions.
- The distinction of two fundamentally different ways of *maximizing partial dialectical interpretations*, viz. the maximization of the actually justified sentences, and the maximization of the interpreted sentences.
- Discussion of the relations between several *types of stages* (or, better, of their non-relations).

Of course some of the above are not entirely new or original, but I claim that the ideas are here at least significantly extended or clarified, given suitable explicitness, or deservedly emphasized.

1.6 Overview of the paper

The paper is structured as follows. Section 2 contains DefLog’s core definition: the dialectical interpretation of theories. In section 3 the notion of dialectical justification is introduced. Dialectical justification is a variant of Dung’s (1995) notions of acceptability and admissibility. In section 4, it is shown how the notion of dialectical justification leads to a characterization of the existence and multiplicity of the dialectical interpretations of theories. In section 5, the focus is on a theory’s stages, i.e., the dialectical interpretations of its parts. In section 6 the relations with Dung’s work (1995) are discussed.

2 DefLog - a logic of dialectical interpretation

The ideas on prima facie justified assumptions can be made formally precise in terms of the logical system DefLog (Verheij 2000a). Its starting point is a simple logical language with two connectives × and ⇝. The first is a unary connective that is used to express the defeat of a statement (a statement’s so-called *dialectical negation*), the latter is a binary connective that is used to express that one statement supports another (*primitive implication*). When φ and ψ are sentences, then $\times\varphi$ expresses that the statement that φ is defeated, and $(\varphi \rightsquigarrow \psi)$ that φ supports ψ, or that ψ follows from φ. Attack, denoted as ⋊, is defined in terms of these two connectives: $\varphi \rtimes \psi$ is defined as $\varphi \rightsquigarrow \times\psi$, and expresses that φ attacks ψ, or that it follows from φ that ψ is defeated. When p, q, r and s are elementary sentences, then $p \rightsquigarrow (q \rightsquigarrow r)$, $p \rightsquigarrow \times(q \rightsquigarrow \times r)$ and $(p \rightsquigarrow q) \rightsquigarrow (p \rightsquigarrow \times(r \rightsquigarrow s))$ are some examples of sentences. For convenience, outer brackets are omitted. Philosophical connotations of the terminology used are here not at issue.

The central definition of DefLog is its notion of the *dialectical interpretation* of a theory. Formally, DefLog’s dialectical interpretations of theories are a variant of Reiter’s (1980) extensions of default theories, Gelfond & Lifschitz's (1988) stable models of logic programming, Dung's (1995) stable extensions of argumentation frameworks, and Bondarenko, Dung, Kowalski & Toni's (1997) stable extensions of assumption-based frameworks.[14]

A theory is any set of sentences, and when it is dialectically interpreted, all sentences in the theory are evaluated, either as justified or as defeated. (This is in contrast with the interpretation of theories in standard logic, where all sentences in an interpreted theory are assigned the same positive value, namely true, e.g., by giving a model of the theory.)

An assignment of the values justified or defeated to the sentences in a theory gives rise to a dialectical interpretation of the theory, when two properties obtain. First, the justified part of the theory must be conflict-free. Second, the justified part of the theory must attack all sentences in the defeated part. Formally the definitions are as follows.

Definition (2.1)

(i) Let T be a set of sentences and φ a sentence. Then T *supports* φ when φ is in T or follows from T by the repeated application of $\leadsto$-Modus ponens (i.e., from $\varphi \leadsto \psi$ and φ, conclude ψ). T *attacks* φ when T supports $\times\varphi$.

(ii) Let T be a set of sentences. Then T is *conflict-free* when there is no sentence φ that is both supported and attacked by T.

(iii) Let Δ be a set of sentences, and let J and D be a partition of Δ, i.e., subsets of Δ that have no elements in common and that have Δ as their union. Then (J, D) *dialectically interprets* the theory Δ when J is conflict-free and attacks all sentences in D. The sentences in J are the *(actually) justified assumptions* of the theory Δ, the sentences in D the *(actually) defeated assumptions*. The sentences in Δ are the theory's *(prima facie justified) assumptions*.

(iv) Let Δ be a set of sentences and let (J, D) dialectically interpret the theory Δ. Then (Supp(J), Att(J)) is a *dialectical interpretation* or *extension* of the theory Δ. Here Supp(J) denotes the set of sentences supported by J, and Att(J) the set of sentences attacked by J. The sentences in Supp(J) are the *justified statements* of the dialectical interpretation, the sentences in Att(J) the *defeated statements*.

Note that when (J, D) dialectically interprets Δ and (Supp(J), Att(J)) is the corresponding dialectical interpretation, J is equal to $\mathrm{Supp}(J) \cap \Delta$, and D to $\mathrm{Att}(J) \cap \Delta$. It is convenient to say that a dialectical interpretation (Supp(J), Att(J)) of a theory Δ *is specified by* J.

Example (2.2): attack and counterattack

(i) Consider the following set of (prima facie justified) assumptions:

$p, q, q \leadsto \times p$

It expresses that the prima facie justified assumption q attacks the prima facie justified assumption p. There is one dialectical interpretation. In it, the assumptions q and $q \leadsto \times p$ are actually justified, and p is defeated. There is one other interpreted sentence, viz. $\times p$, that is justified. Formally, this example is equal to the presumption of innocence example above (section 1.4).

(ii) Consider the following set of assumptions:

$p, q, q \leadsto \times p, r, r \leadsto \times q$

The attack of q by the prima facie justified assumption r has been added to the assumptions of the previous example. There is one dialectical interpretation. In it, the assumptions p, $q \leadsto \times p$, r and $r \leadsto \times q$ are actually justified, and q defeated. There is one other interpreted sentence, viz. $\times q$, that is justified.

[14] In section 6, a formal connection with Dung's (1995) work is discussed. Verheij (2000a) gives other relations between DefLog and related formalisms. See also Dung (1995) for relations of his work with other formalisms. To guide intuition, the following may be useful. A default p : q / r (as in Reiter's 1980) would in DefLog be translated to two conditionals, viz. $p \leadsto r$ and $\neg q \leadsto \times(p \leadsto r)$. The second says that the former is defeated in case of $\neg q$. This corresponds to the intuition underlying the default that r follows from p as long as q can consistently be assumed. (Note however that the properties of ordinary negation $\neg$ are not part of DefLog proper.) A rule in logic programming $p \leftarrow q, {\sim}r$ corresponds in DefLog to two conditionals, viz. $q \leadsto p$ and $r \leadsto \times(q \leadsto p)$. The second says that $q \leadsto p$ is defeated in case of r. This corresponds to the intuition underlying the program rule that p follows from q when r is not provable, but that p does not follow from q when r is provable.

This example of attack and counterattack shows the phenomenon of *reinstatement*, that is typical for defeasible reasoning: an assumption that is defeated can become justified when there is additional information.

Example (2.3): dialectical negation and double dialectical negation

(i) Consider the following set of assumptions:

$p, \times p$

It expresses that it is prima facie justified that the prima facie justified assumption p is defeated. There is one dialectical interpretation. In it, the assumption $\times p$ is actually justified, and p is defeated. There is no other interpreted sentence.

(ii) Consider the following set of assumptions:

$p, \times p, \times\times p$

It adds to the previous example that it is prima facie justified that p’s defeat is defeated. There is one dialectical interpretation. In it, the assumptions p and $\times\times p$ are actually justified, and $\times p$ is defeated. There is no other interpreted sentence.

The examples of dialectical negation and double dialectical negation show the asymmetry between a sentence and its dialectical negation: whereas the fact that an assumption’s dialectical negation is justified indicates the assumption’s defeat, an assumption’s being justified does not indicate its dialectical negation’s defeat. (See also section 12.2 of Verheij (2000a) on symmetric DefLog.) Note also that neither of the double negation rules of standard logic hold (implying that dialectical negation differs from classical negation and from intuitionistic negation): when p is justified, $\times\times p$ is not necessarily also justified, nor is p necessarily justified, when $\times\times p$ is. The former is shown by the first of the above two examples, the latter follows by considering the single assumption $\times\times p$. In its unique dialectical interpretation, only $\times\times p$ is justified and only $\times p$ defeated.

Example (2.4): primitive implication

(i) Consider the following set of assumptions:

$p \leadsto q, q \leadsto (r \leadsto s), p, r$

It expresses that it is prima facie justified that q follows from the prima facie justified assumption p, and that it follows from q that s follows from the assumption r. There is one dialectical interpretation, in which all assumptions are justified. There are three additional interpreted sentences, viz. q, $r \leadsto s$ and s, that are all justified.

(ii) Consider the following set of assumptions:

$p \leadsto q, q \leadsto (r \leadsto s), p, r, t \leadsto \times(p \leadsto q), t$

The prima facie justified assumption t, attacking the assumption that q follows from p, has been added to the previous example. In the unique dialectical interpretation of these assumptions, all assumptions are justified, except for $p \leadsto q$ that is defeated. There is one other interpreted sentence, viz. $\times(p \leadsto q)$ that is justified. Note that q, $r \leadsto s$ and s are neither justified nor defeated.

(iii) Consider the following set of assumptions:

$p \leadsto q, \times q$

There is one dialectical interpretation, in which both assumptions are justified. The sentence q is defeated, and no other sentence is interpreted. This shows that contraposition (From a conditional and the negation of its consequent, conclude its negated antecedent) does not hold.

The examples of the primitive implication also show that the analogues of the paradoxes of the material implication (viz. that $\psi \rightarrow (\varphi \rightarrow \psi)$ and $\neg\varphi \rightarrow (\varphi \rightarrow \psi)$ are logically true) do not hold: when $\times q$ is justified, no conditional with q as its antecedent necessarily follows (example (iii)), and when p is justified, no conditional with p as its consequent necessarily follows (example (i)).

Many theories have a unique dialectical interpretation. For instance, a conflict-free theory always has a unique dialectical interpretation, namely the dialectical interpretation specified by the theory itself. Examples of theories with no or with several dialectical interpretations are the following:

Example (2.5): loops of attacks

(i) Consider the following set of assumptions:

$p, p \leadsto \times p$

It expresses that p attacks itself. The assumptions have no dialectical interpretation.

(ii) Consider the following set of assumptions:

$p, q, p \rightsquigarrow \times q, q \rightsquigarrow \times p$

It expresses that p attacks q and vice versa. The assumptions have two dialectical interpretations. In one, all assumptions are justified, except for p that is defeated. In the other, only q is defeated.

(iii) Consider the following set of assumptions:

$p, q, r, p \rightsquigarrow \times q, q \rightsquigarrow \times r, r \rightsquigarrow \times p$

It expresses that p attacks q, which attacks r, which on its turn attacks p. The assumptions have no dialectical interpretation.

Example (2.6): the Nixon diamond

(i) Consider the following set of assumptions:

$q, r, q \rightsquigarrow p, r \rightsquigarrow \times p$

It expresses that q supports p and that r attacks p. These assumptions have no dialectical interpretation.

(ii) It may be thought that the assumptions of the previous example are a formalization of the so-called Nixon diamond, a famous example in nonmonotonic logic. In fact, the previous example is *not* the analogue in DEFLOG of the Nixon diamond. In Reiter's (1980) default logic, it looks thus:

$q; r; q : p / p; r : \neg p / \neg p$

These express that Nixon is a quaker and a republican, and that quakers are pacifists, while republicans are non-pacifists. Reiter's definitions give rise to two extensions. In one, p follows by the application of the first default, in the other, ¬p follows by the application of the second default. A representation of the Nixon-diamond in DEFLOG[15] takes the following assumptions:

$q, r, q \rightsquigarrow p, r \rightsquigarrow \text{not-}p, p \rightsquigarrow \times(r \rightsquigarrow \text{not-}p), \text{not-}p \rightsquigarrow \times(q \rightsquigarrow p)$

The latter two conditionals express that when p is actually justified, it is defeated that r implies not-p, and that when not-p is actually justified, it is defeated that q implies p. In each of the two dialectical interpretations of these assumptions, one of $q \rightsquigarrow p$ and $r \rightsquigarrow \text{not-}p$ is defeated, the other justified. In the DEFLOG formalization, the conditionals $q \rightsquigarrow p$ and $r \rightsquigarrow \text{not-}p$ stand for the 'application' of the Reiter defaults, while the conditionals $p \rightsquigarrow \times(r \rightsquigarrow \text{not-}p)$ and $\text{not-}p \rightsquigarrow \times(q \rightsquigarrow p)$ express when that application is blocked. The difference between the Reiter and the DEFLOG formalization has to do with the fact that Reiter's defaults are inconclusive (their consequent does not always follow when their antecedent obtains), while DEFLOG uses conditionals that are prima facie justified, just like other assumptions. Cf. also Verheij (2000a), section 11.2. Note that the opposition between p and not-p is not represented in terms of dialectical negation (i.e., as p and $\times p$), showing that it can make sense that standard and dialectical negation are used side-by-side, each for its own purpose.

Example (2.7)

(i) The three theories $\{p, p \rightsquigarrow \times p\}$, $\{p, p \rightsquigarrow q, \times q\}$ and $\{p_i \mid i \text{ is a natural number}\} \cup \{p_j \rightsquigarrow \times p_i \mid i \text{ and } j \text{ are natural numbers, such that } i < j\}$ lack dialectical interpretations, each in a different way. The first is a simple attack loop. The second shows that the defeat of a supported statement requires more than just assuming its defeat. The third theory is more complex. It can be seen as follows that it lacks a dialectical interpretation. Assume that there is a dialectical interpretation E in which for some natural number n p_n is justified. Then all p_m with $m > n$ must be defeated in E, for if such a p_m were justified, p_n could not be justified. But that is impossible, for the defeat of a p_m with $m > n$ can only be the result of an attack by a justified $p_{m'}$ with $m' > m$. As a result, no p_i can be justified in E. But then all p_i must be defeated in E, which is impossible since the defeat of a p_i can only be the result of an attack by a justified p_j with $j > i$. (Note that any finite subset of the latter theory has a dialectical interpretation, while the whole theory does not.)

(ii) The three theories $\{p, q, p \rightsquigarrow \times q, q \rightsquigarrow \times p\}$, $\{p_i, p_{i+1} \rightsquigarrow \times p_i \mid i \text{ is a natural number}\}$ and $\{\times^i p \mid i \text{ is a natural number}\}$ have two dialectical interpretations. Here $\times^i p$ denotes, for any natural number i, the sentence composed of a length i sequence of the connective ×, followed by the constant p. (Note that each finite subset of the latter theory has a unique dialectical interpretation.)

[15] Verheij (2000a) and (2002) give different analyses of the Nixon diamond. The former contains an error, the latter is unnecessarily entangled and is farther away from Reiter's default logic than the analysis given here. The present analysis follows the translation of Reiter's defaults to DEFLOG sentences as discussed by Verheij (2000a).

DefLog's connectives ⇝ and × are obviously reminiscent of propositional logic's connectives → and ¬. Also some of DefLog's definitions remind of propositional logic. These likenesses have been incorporated on purpose. In fact, DefLog has been carefully designed to be as close as possible to propositional logic (as the paradigmatic example of deductive logic), while retaining the essence of defeasible logic.

The examples already showed some differences between DefLog's connectives and those of propositional logic. Another difference is that the set $\{p, \times p\}$ is not 'inconsistent' or 'unsatisfiable', from the dialectical point of view: the theory $\{p, \times p\}$ has a unique dialectical interpretation in which p is defeated and $\times p$ justified. Of course $\{p, \neg p\}$ is classically inconsistent. The theory $\{p, \times p\}$ shows the essence of dialectical negation: the dialectical negation of a sentence in a sense 'prevails' over the sentence. By this prevalence of dialectical negation, assumptions are only prima facie justified: a prima facie assumption is not actually justified when the dialectical negation of the assumption is (actually) justified.

Note that the prevalence relation between a sentence p and its weak negation ~p in logic programming is exactly opposite to that between a sentence p and its dialectical negation $\times p$: in logic programming ~p can be assumed as long as p is not provable, while in dialectical argumentation p can be assumed as long as $\times p$ is not justified.

The theory $\{p, \times p\}$ also shows that dialectical interpretation is not simply maximal consistency: whereas the maximal consistent subset $\{\times p\}$ corresponds to a (the) dialectical interpretation, $\{p\}$ does not. Verheij (2000a) gives much more information on DefLog, for instance, on different ways to adapt DefLog to incorporate the classical logical connectives. Once again: DefLog's connectives ⇝ and × are not meant to replace the classical connectives; they express different concepts.

It is not hard to see that DefLog is non-monotonic, for instance in the following sense: when a sentence is justified in some dialectical interpretation of a theory, it need not be in a dialectical interpretation of a larger theory. The simplest example is provided by the theories $\{p\}$ and $\{p, \times p\}$. Both have only one dialectical interpretation. In the dialectical interpretation of $\{p\}$, p is justified, but in that of $\{p, \times p\}$, p is defeated (and $\times p$ justified).

Notwithstanding the simple structure of DefLog's logical language (with only two connectives, viz. ⇝ and ×), many central notions of dialectical argumentation can be analyzed in terms of it. For instance, it is possible to define an inconclusive conditional (i.e., a conditional of which the consequent does not always follow when its antecedent obtains) in terms of DefLog's defeasible conditional (that is defeasible in the same way as any other statement). DefLog's expressiveness also allows an integrated analysis of Toulmin's (1958) warrants and backings and Pollock's (1987) undercutting and rebutting defeaters. A warrant and an undercutter can be seen as the support and attack, respectively, of the relation between a reason and its conclusion. In DefLog this can be expressed by sentences of the forms $\varphi \rightsquigarrow (\psi \rightsquigarrow \chi)$ and $\varphi \rightsquigarrow \times(\psi \rightsquigarrow \chi)$. Whereas Pollock treats undercutting and rebutting defeaters as separate concepts, in DefLog it is natural to consider both as different instances of the general phenomenon of defeat. Cf. Verheij (2000a, 2001a).

3 Dialectically justifying arguments

Before we proceed to the notion of dialectical justification, some terminology needs to be introduced.

Definition (3.1)

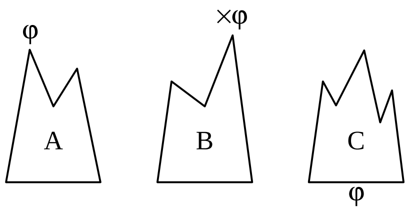
(i) A set of sentences is an *argument* when it is conflict-free. If Δ is a set of sentences, a Δ-*argument* is an argument that is a subset of Δ.

(ii) Let φ be a sentence. An argument C is an *argument for* φ if C supports φ. An argument C is an *argument against* φ if C attacks φ. The sentences in an argument C are also called its *premises*, the sentences φ such that C supports φ, its *conclusions*.

(iii) An argument C *attacks* an argument C' if C attacks a sentence in C'.

(iv) Arguments C and C' are *compatible* when $C \cup C'$ is an argument, and otherwise *incompatible*. The arguments in a collection $\{C_i\}_{i \in I}$ are *compatible* if their union $\cup_{i \in I} C_i$ is an argument, otherwise *incompatible*.

In the following figure, three arguments are graphically suggested.



The bottoms of the alpine shapes consist of the premises of the argument; the tops are the conclusions. Argument A has conclusion φ, argument B conclusion ×φ and argument C has premise φ. B attacks C, but not necessarily A (since φ might not be a premise of A). A and B are incompatible, and B and C are too.

When a theory has a dialectical interpretation, the set of sentences of the theory that are justified in the interpretation are clearly an argument. It has a special property:

Proposition (3.2)

Let E be a dialectical interpretation of a theory Δ. Then J(E) ∩ Δ is a Δ-argument that attacks any Δ-argument C that is incompatible with J(E) ∩ Δ. Here J(E) denotes the set of justified statements of the dialectical interpretation E.

Proof: Since E is a dialectical interpretation, J(E) ∩ Δ is conflict-free. Hence a Δ-argument C that is incompatible with J(E) ∩ Δ cannot be a subset of J(E) ∩ Δ. Therefore there is a sentence φ in C that is not in J(E) ∩ Δ. Since E is a dialectical interpretation, it is in D(E), the set of defeated sentences of the dialectical interpretation E. But for any sentence φ in D(E) it holds by the definition of dialectical interpretations that J(E) ∩ Δ attacks φ. In other words, J(E) ∩ Δ attacks C.

Arguments with the property that J(E) ∩ Δ has in proposition (3.2) are said to be dialectically justifying:

Definition (3.3)

(i) A Δ-argument C is *dialectically justifying* with respect to Δ if and only if C attacks every Δ-argument C’ that is incompatible with C.

(ii) A sentence φ is *dialectically justifiable* with respect to a set of sentences Δ if and only if there is a Δ-argument C for φ that is dialectically justifying with respect to Δ. Such an argument C is then called a *dialectical justification of* φ, and C *dialectically justifies* φ with respect to Δ. A sentence φ is *dialectically defeasible* with respect to Δ if and only if ×φ is dialectically justifiable with respect to Δ. If C is a dialectical justification of φ, then the argument C *dialectically defeats* φ with respect to Δ.

(iii) A sentence φ is *dialectically interpretable* with respect to a set of sentences Δ if and only if it is dialectically justifiable or dialectically defeasible with respect to Δ. A sentence φ is *dialectically ambiguous* with respect to a set of sentences Δ if and only if it is both dialectically justifiable and dialectically defeasible with respect to Δ.

The argument {p, r, r → ×q} dialectically justifies p with respect to the theory {p, q, r, q → ×p, r → ×q}. The argument {p} does not dialectically justify p since the incompatible argument {q, q → ×p} is not attacked. The argument {r, r → ×q} dialectically defeats q with respect to the theory.

The sentences p and q are dialectically ambiguous with respect to the theory {p, q, p → ×q, q → ×p} since the argument {p, p → ×q} dialectically justifies p and dialectically defeats q, and likewise for q.

The sentence p is not dialectically interpretable with respect to the theory {p, p → ×p}.

When an argument is dialectically justifying with respect to a theory, it dialectically justifies all the sentences it supports.

Note the similarity of dialectical justification with Dung’s (1995) admissibility. Whereas admissibility requires that there is an attack against each attack, dialectical justification requires something stronger: there must be an attack against each incompatibility. See section 6 for a further discussion of the relations between the two notions.

4 The existence and multiplicity of dialectical interpretations

When a theory has a dialectical interpretation, all sentences in the theory are dialectically interpretable. In other words, dialectical justification is a kind of 'local' dialectical interpretation. This is an immediate corollary of proposition (3.2):

Corollary (4.1)

Let E be a dialectical interpretation of the theory Δ. Then all sentences in the theory are dialectically justifiable or dialectically defeasible with respect to Δ.

Proof: By proposition (3.2), $J(E) \cap \Delta$ dialectically justifies or defeats all sentences in Δ.

Note that corollary (4.1) gives a necessary condition for the existence of a dialectical interpretation: when there is a sentence in a theory that is not dialectically interpretable, there cannot be a dialectical interpretation. Corollary (4.1) can explain all examples of theories without dialectical interpretations that have been encountered: in all, there is a sentence that is not dialectically interpretable. Nevertheless the condition in the corollary is *not* sufficient for the existence of a dialectical interpretation, as the theory $\Delta = \{p, q, p \dashrightarrow \times q, q \dashrightarrow \times p, r, r \dashrightarrow \times r, s, s \dashrightarrow \times s, p \dashrightarrow \times r, q \dashrightarrow \times s\}$ shows. It has no dialectical interpretation. Nevertheless all sentences in the theory are dialectically justifiable or defeasible with respect to Δ. The Δ-argument $\{p, p \dashrightarrow \times q, p \dashrightarrow \times r\}$ dialectically justifies p and dialectically defeats q and r, while $\{q, q \dashrightarrow \times p, q \dashrightarrow \times s\}$ dialectically justifies q and dialectically defeats p and r.

The notion of dialectical justification plays the central role in theorem (4.3) below, that shows exactly under which circumstances a theory has a dialectical interpretation. One additional definition is needed.

Definition (4.2)

Let C be an argument. A sentence φ is *dialectically justifiable in the context* C with respect to a theory Δ if it is supported by a dialectically justifying argument of the theory that contains C, and *dialectically defeasible in the context* C if $\times\varphi$ is supported by a dialectically justifying argument that contains C.

Now the theorem can be formulated:

Theorem (4.3)

A theory Δ has a dialectical interpretation if and only if there is an argument C in the context of which all sentences in Δ are either dialectically justifiable or dialectically defeasible with respect to the theory, but not both.

(The proof follows below.) In other words, a theory has a dialectical interpretation if and only if there is a context in which all sentences of the theory are dialectically interpretable, while none is dialectically ambiguous. Theorem (4.3) is closely related to corollary (4.1) above that says that the dialectical interpretability of all sentences of a theory is necessary for the existence of a dialectical interpretation. Theorem (4.3) says that the dialectical interpretability of all sentences *in a context with no dialectical ambiguities* is both necessary and sufficient for the existence of a dialectical interpretation. In other words, after fixing all choices allowed by dialectically ambiguous sentences in the theory, it suffices for the existence of a dialectical interpretation that all sentences in the theory are either dialectically justifiable or dialectically defeasible. The example that showed why the dialectical interpretability of all sentences of a theory is not sufficient for the existence of a dialectical interpretation, shows what can go wrong: the dialectical justification of one sentence (or its dialectical negation) need not be compatible with that of another when there is a dialectical ambiguity. In other words, the dialectical justification of sentences can depend on the particular choice allowed by a dialectical ambiguity. Dialectical justifications that require different choices cannot be 'glued' to form a dialectical interpretation.

Three properties of dialectical justification are essential in the proof of the theorem (4.3):

Proposition (4.4)

(i) *Localization:* Let E be a dialectical interpretation of a theory Δ. Then there is a collection $\{C_i\}_{i \in I}$ of arguments that covers $J(E) \cap \Delta$ (i.e., $J(E) \cap \Delta$ is equal to $\cup_{i \in I} C_i$), that are dialectically justifying with respect to the theory.

(ii) *Union:* If C and C' are compatible arguments, that are dialectically justifying with respect to a theory Δ, then also $C \cup C'$ is dialectically justifying with respect to the theory. (Similarly, for collections of dialectically justifying arguments: the union of a compatible collection of dialectically justifying arguments is again dialectically justifying.)

(iii) *Separation at the base:* If C and C' are incompatible arguments, that are dialectically justifying with respect to a theory Δ, then there is a sentence in Δ that is both dialectically justifiable and defeasible with respect to Δ. (Similarly, for collections of dialectically justifying arguments: given an incompatible collection of dialectically justifying arguments, there is a sentence in the theory that is both dialectically justifiable and defeasible.)[16]

Proof: Localization follows from corollary (4.1): it shows that $J(E) \cap \Delta$ is itself dialectically justifying with respect to Δ. The union property (for pairs of arguments) is seen as follows. Let C and C' be compatible dialectically justifying arguments, and let the argument C'' be incompatible with $C \cup C'$. Assume first that C'' is incompatible with C. Then clearly C attacks C''. Assume second that C'' is compatible with C. Then C' is incompatible with the argument $C \cup C''$, and therefore attacks it. Since C and C' are compatible, it then follows that C' attacks C''. The proof of the general case of the union property requires some extra care, but is similar. The property of separation at the base follows directly from the definition of dialectical justification: when C and C' are dialectically justifying and incompatible, they attack each other. Then there is a sentence in each (and therefore in the theory itself) that is attacked by the other. The general case of the separation property can be reduced to the case of pairs of arguments.

Proof of theorem (4.3): First let E be a dialectical interpretation of Δ. Then by the localization property $J(E) \cap \Delta$ can be covered by arguments that are dialectically justifying with respect to Δ. By the union property, it then follows that $J(E) \cap \Delta$ is also dialectically justifying. (In fact, the proof of corollary (4.1) at the beginning of the section directly shows that $J(E) \cap \Delta$ is dialectically justifying.) As a result, $J(E) \cap \Delta$ is a context as in the theorem since by the fact that $J(E) \cap \Delta$ is dialectically justifying and by the definition of dialectical interpretations all sentences in Δ are dialectically interpretable in the context of $J(E) \cap \Delta$, and since by the fact that $J(E) \cap \Delta$ is conflict-free there is no dialectically ambiguous sentence in that context. Second let C be a context as in the theorem, and let, for all sentences φ, C_φ be a Δ-argument dialectically justifying or defeating φ in the context C. The collection of the C_φ is compatible since by the property of separation at the base there would otherwise be a sentence in the theory that is dialectically ambiguous in the context C. By the union property, the union of the C_φ is dialectically justifying. It specifies a dialectical interpretation of Δ.

The proof shows that dialectical interpretations can be built by 'gluing' dialectically justifying arguments. This suggests that a (set-theoretically minimal) argument that dialectically justifies a sentence, is a kind of dialectical proof of the sentence. Similarly, such a dialectical proof of the dialectical negation of a sentence is a kind of dialectical refutation of the sentence.

The following theorem provides a general answer to the problems of the existence and multiplicity of dialectical interpretations. It is a corollary of theorem (4.3) above:

Theorem (4.5)

Let *n* be a natural (or cardinal) number (possibly 0). A theory Δ has exactly *n* dialectical interpretations if and only if *n* is equal to the maximal number of mutually incompatible arguments C in the context of which all sentences in Δ are either dialectically justifiable or dialectically defeasible with respect to the theory, but not both.

5 Stages

Even if a theory has no dialectical interpretation, its subsets can have dialectical interpretations. The dialectical interpretations of subsets of a theory are called the theory's *stages*.[17] The dialectical interpretations of the subsets of a theory can be regarded as preliminary stages on the path towards a dialectical interpretation of the whole theory. One could say that at these preliminary stages less

[16] The property is called separation *at the base* since the dialectically ambiguous sentence can be found in the theory itself.

[17] For the development of my ideas on stages, see also Verheij 1996a and 1996b.

information as it is contained in the theory, is taken into account than at a dialectical interpretation. Even if the theory as a whole lacks a dialectical interpretation, its stages can provide interesting information about the theory.

In section 5.1, stages are defined. Section 5.2 discusses a special class of stages, viz. those that are specified by a dialectically justifying argument. In section 5.3, it is shown that surprisingly there are few relations between the different types of stages.

5.1 Definition

A theory's stages are the dialectical interpretations of subsets of the theory.

Definition (5.1): stages

Let Δ, J and D be sets of sentences. Then (J, D) is a *stage of the theory* Δ if and only if it is a dialectical interpretation of a subset of Δ. The set $\Delta \cap (J \cup D)$ is the *scope* of the stage. The sets $J \cap \Delta$ and $D \cap \Delta$ are the *j-scope* and the *d-scope* of the stage, respectively. A sentence φ in Δ that is in the scope of a stage S is *taken into account at the stage* S.

For instance, the stages of the theory $\{p, q, q \rightsquigarrow \times p\}$ are specified by the sets $\emptyset$, $\{p\}$, $\{q\}$, $\{q \rightsquigarrow \times p\}$, $\{p, q\}$, $\{p, q \rightsquigarrow \times p\}$ and $\{q, q \rightsquigarrow \times p\}$. The latter has the whole theory as its scope (the assumption p is defeated) and specifies the theory's unique extension. The stages of the theory $\{p, \times p, p \rightsquigarrow q\}$ are specified by the sets $\emptyset$, $\{p\}$, $\{\times p\}$, $\{p \rightsquigarrow q\}$, $\{p, p \rightsquigarrow q\}$ and $\{\times p, p \rightsquigarrow q\}$. The theory's unique dialectical interpretation is specified by $\{\times p, p \rightsquigarrow q\}$. Note that the scopes of the stages specified by $\{\times p\}$ and $\{\times p, p \rightsquigarrow q\}$ include the sentence p.

Not all subsets of a theory occur as the scope of one of the theory's stages and the stages of a theory correspond exactly to the interpretations that are specified by the conflict-free subsets of Δ. (Cf. Verheij 2000a.) Note that *extensionally* stages coincide with conflict-free subsets, but not *intensionally*. One notion associated with the stages of a theory is their scope, i.e., the part of the theory that has been taken into account, which is not as readily associated with the theory's conflict-free subsets.

A stage is said to succeed another when it takes more assumptions of the theory into account. For instance, with respect to the theory $\{p, q, q \rightsquigarrow \times p\}$, the stage specified by $\{p\}$ precedes the stage specified by $\{q, q \rightsquigarrow \times p\}$. The latter is in fact the theory's dialectical interpretation. In the former only p is taken into account, in the latter q is also taken into account as an attack of p.

The idea of stages succeeding each other gives rise to two partial orders on the set of stages of a theory: one in terms of the theory's assumptions that are justified in the stage (the j-scope), the other in terms of the theory's assumptions that are taken into account at the stage (the scope).

Definition (5.2): a stage preceding or succeeding another stage

Let S and S' be stages of the theory Δ. S *precedes* S', denoted as $S \prec_\Delta S'$, if the scope of S is a proper subset of that of S'. S *compatibly precedes* S', denoted as $S \sqsubset_\Delta S'$, if the j-scope of S is a proper subset of the j-scope of S'. A stage S' *(compatibly) succeeds* a stage S when S (compatibly) precedes S'.

Note that $\prec_\Delta$ and $\sqsubset_\Delta$ are not defined in terms of set inclusion of the extents of stages (i.e., all interpreted sentences), but of their scopes (i.e., the interpreted assumptions). They are strict partial orders on the set of stages of a theory.

It is a natural step to accentuate the stages that are maximal with respect to the two partial orders, as follows.

Definition (5.3): compatibility classes[18] *and maximal stages*

A stage S is a *compatibility class of the theory* Δ if and only if it has maximal j-scope among Δ's stages. A stage S is a *maximal stage of the theory* Δ if and only if it has maximal scope among the stages of Δ.

Compatibility classes correspond to the maximal conflict-free sets of the theory. When a stage is a compatibility class of a theory, taking additional sentences into account requires the defeat of a sentence

[18] Verheij (2000a) speaks of satisfiability classes instead of compatibility classes.

that is justified in the stage: its j-scope is not contained in the j-scope of any successor stage. Maximal stages are stages in which no additional sentences of the theory can be taken into account: a maximal stage has no successors.

Every theory Δ has one or more compatibility classes, but there exist theories that have no maximal stage. Since every finite theory has a maximal stage (but not necessarily a dialectical interpretation), an example must be infinite:

Example (5.4): a theory without maximal stage

The theory $\Delta = \{p_i \mid i$ is a natural number$\} \cup \{p_j \leadsto \times p_i \mid i$ and j are natural numbers, such that $i < j\}$ has no maximal stage. This can be seen as follows. Among its stages are the interpretations S_n specified by the sets $\{p_n\} \cup \{p_j \leadsto \times p_i \mid i$ and j are natural numbers, such that $i < j\}$, where n is a natural number. In a stage S_n, p_n is justified and every p_i with $i < n$ is defeated. When $i < j$, the scope of S_i is a subset of the scope of S_j. Moreover, the scopes of the stages S_n exhaust the whole theory. As a result, a maximal stage must have the whole theory as its scope, i.e., must be a dialectical interpretation. However, the theory does not have a dialectical interpretation, cf. example (2.7), part (i). Note that every finite subset of Δ has a dialectical interpretation.

Dialectical interpretations are maximal stages, but not in general vice versa. If a theory has a dialectical interpretation, then every maximal stage of the theory is a dialectical interpretation. Every maximal stage of a theory is a compatibility class, but not in general vice versa. Every dialectical interpretation of a theory is a compatibility class, but not in general vice versa.

Different satisfiability classes, different maximal stages and different dialectical interpretations are incompatible, i.e., have incompatible sets of justified sentences.

5.2 *Dialectically justified stages*

Among all stages, a special group can be distinguished: the dialectically justified stages, i.e., the stages that are specified by a dialectically justifying argument. The definition is as follows.

Definition (5.5): dialectically justified stages

A stage S is a *dialectically justified stage of the theory* Δ if and only if S is a stage of Δ, for which it obtains that J(S) dialectically justifies every sentence in J(S) and dialectically defeats every sentence in D(S).

A dialectically justified stage that has no compatible successor is a dialectically preferred stage (an analog of Dung's 1995 preferred extensions), and a dialectically justified stage without successor is a maximal dialectically preferred stage:

Definition (5.6): dialectically preferred stages and maximal dialectically preferred stages

A stage S is a *dialectically preferred stage of the theory* Δ if and only if it has maximal j-scope among the dialectically justified stages of Δ. A stage S is a *maximal dialectically preferred stage of the theory* Δ if and only if it has maximal scope among the dialectically justified stages of Δ.

The dialectically preferred stages of a theory are the $\sqsubset$-maximal elements among the theory's dialectically justified stages. A theory's maximal dialectically preferred stages the $\prec$-maximal elements.

Different dialectically preferred stages and different maximal dialectically preferred stages are not compatible.

Maximal dialectically preferred stages are of course dialectically preferred stages and dialectical interpretations are maximal dialectically preferred stages. However, there exist theories that have dialectically preferred stages that are not maximal dialectically preferred stages, and theories that have maximal dialectically preferred stages that are not dialectical interpretations. The following is an example of a theory with a dialectically preferred stage that is not maximal dialectically preferred.

Example (5.7)

The theory $\{p, q, r, p \leadsto \times q, q \leadsto \times p, q \leadsto \times r, r \leadsto \times r\}$ has the stage specified by $\{q, p \leadsto \times q, q \leadsto \times p, q \leadsto \times r, r \leadsto \times r\}$, in which p and r are defeated and q is justified, as dialectical interpretation, and therefore as maximal dialectically preferred stage. The stage specified by $\{p, p \leadsto \times q, q \leadsto \times p, q \leadsto \times r\}$, in which p is justified, q is defeated and r is not taken into account, is a dialectically preferred stage, that is not maximal dialectically preferred.

Dialectically preferred stages are not necessarily compatibility classes and compatibility classes are not necessarily dialectically preferred stages.

Every theory has a dialectically preferred stage, but not all theories have a maximal dialectically preferred stage. A theory without a maximal dialectically preferred stage must be infinite however. The construction of an example of a theory without a maximal dialectically preferred stage is rather involved:

Example (5.8): a theory without maximal dialectically preferred stage

Consider the theory Δ consisting of the following sentences:

p_i, q_i, r_i, for every natural number i
$p_i \rtimes p_j$ for all i and j with $i < j$
$p_i \rtimes q_i$ and $q_i \rtimes p_i$ for all i
$p_i \rtimes r_k$ for all i and k with $k \leq i$
$r_k \rtimes r_k$ for all k

Then the following are the ‘initials’ of some of Δ’s stages:

S_0:	p_0 (q_0) (r_0)	(p_1) q_1 -	(p_2) q_2 -	(p_3) q_3 -	(p_4) q_4 -	...
S_1:	(p_0) q_0 (r_0)	p_1 (q_1) (r_1)	(p_2) q_2 -	(p_3) q_3 -	(p_4) q_4 -	...
S_2:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	p_2 (q_2) (r_2)	(p_3) q_3 -	(p_4) q_4 -	...
S_3:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	(p_2) q_2 (r_2)	p_3 (q_3) (r_3)	(p_4) q_4 -	...
S_4:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	(p_2) q_2 (r_2)	(p_3) q_3 (r_3)	p_4 (q_4) (r_4)	...
...	...	...	...	...	...	...

The sentences in brackets ( ) are defeated at the stage. The other listed sentences are justified. The hyphens - indicate sentences that are not taken into account. For instance, at S_0, p_0 is justified, q_0 is defeated and r_1 is not taken into account. For every natural number i, S_i is defined as the stage at which

(i) p_i is justified and, for every j such that $i \neq j$, p_j is defeated, and
(ii) q_i is defeated and, for every j such that $i \neq j$, q_j is justified, and
(iii) for every j such that $i \geq j$, r_i is defeated and, for every j such that $i < j$, r_j is not taken into account, and
(iv) every sentence in Δ of the form $\varphi \rtimes \psi$ is justified.

The following properties obtain, as is proven below:

a. Each stage S_i is dialectically preferred.
b. S_i and S_j are incompatible if $i \neq j$.
c. If $i < j$, then the scope of S_i is a proper subset of the scope of S_j.
d. If a stage S is dialectically preferred, such that, for some i, p_i is justified in S, then S is equal to S_i.
e. If a stage S is dialectically preferred, such that no p_i is justified, then all p_i are defeated, all q_i are justified and no r_i is taken into account in S. The scope of this stage is properly contained in the scope of all of the stages S_i.
f. Δ has no maximal dialectically preferred stage.

Proof: The proof of the properties is given by Verheij (2000a).

Note that though every finite stage of the sample theory has a maximal dialectically preferred stage, the whole theory does not.

5.3 The relations between the types of stages[19]

Several special kinds of stages of a theory have been discussed. Apart from a theory’s dialectical interpretations (also called extensions), we encountered maximal stages and compatibility classes (in section 5.1) and dialectically preferred stages and maximal dialectically preferred stages (in section 5.2). One might hope that there are close relations between the types of stages. For instance, the property ‘The maximal stages of a theory coincide with its dialectically preferred stages’ would be most welcome. It would show a connection between two ideas: on the one hand the idea that it is relevant to take as many sentences of a theory into account as possible (cf. the maximal stages), on the other the idea that it is

[19] This section extends my earlier work on the relations between types of stages (Verheij 1996a).

relevant to defend against attack and incompatibility (cf. the dialectically preferred stages). Unfortunately, that property does not hold.[20]

The relations between the types of stages of section 5.1 (the compatibility types) and the relations between those in section 5.2 (the dialectical justification types) have already been investigated. If $\underline{E}$, $\underline{M}$, $\underline{CC}$, $\underline{S}$, $\underline{DJ}$, $\underline{P}$ and $\underline{MP}$ denote the sets of dialectical interpretations (extensions), maximal stages, compatibility classes, stages, dialectically justified stages, dialectically preferred stages and maximal dialectically preferred stages of a theory, respectively, the relations between the types within the same group can be summarized as in the following figure.

$$\begin{tikzcd} \textit{compatibility types} & \underline{E} \arrow[r, hook] & \underline{M} \arrow[r, hook] & \underline{CC} \arrow[r, hook] & \underline{S} \\ \textit{dialectical justification types} & \underline{E} \arrow[r, hook] & \underline{MP} \arrow[r, hook] & \underline{P} \arrow[r, hook] & \underline{DJ} \end{tikzcd}$$

The arrows indicate inclusion maps between the sets of stages.

No other interesting inclusion relation between the types of stages exists. Surprisingly, it is not even the case that the compatible successor stages of a dialectically preferred stage or a maximal dialectically preferred stage always include a compatibility class or a maximal stage. This can be seen by considering the dialectically justified restrictions of stages, i.e., the largest substage of a stage that is dialectically justified. It is not hard to show that such dialectically justified restrictions exist. Cf. Verheij (2000a). It can be shown that the images of the sets $\underline{M}$ and $\underline{CC}$ under the restriction map are not in general included in $\underline{MP}$ or $\underline{P}$, and that the originals of $\underline{MP}$ and $\underline{P}$ do not in general include $\underline{M}$ or $\underline{CC}$.

Example (5.9)

(i) The theory $\{p, q, r, p \rightsquigarrow \times q, q \rightsquigarrow \times r, r \rightsquigarrow \times r\}$ has four maximal stages (but no dialectical interpretation). One of them is M_1 specified by $\{p, p \rightsquigarrow \times q, q \rightsquigarrow \times r, r \rightsquigarrow \times r\}$: p is justified, q is defeated and r is not taken into account. M_1's dialectically justified restriction is the theory's maximal dialectically preferred stage, specified by $\{p, p \rightsquigarrow \times q, q \rightsquigarrow \times r\}$. Another maximal stage is M_2, specified by $\{q, p \rightsquigarrow \times q, q \rightsquigarrow \times r, r \rightsquigarrow \times r\}$: p is not taken into account, q is justified and r is defeated. The dialectically justified restriction of M_2 is the empty stage, which is not dialectically preferred. M_2 is a maximal stage, that is not the dialectically justified restriction of a dialectically preferred stage. (The other two maximal stages are specified by $\{p, q, q \rightsquigarrow \times r, r \rightsquigarrow \times r\}$ and $\{p, r, p \rightsquigarrow \times q, q \rightsquigarrow \times r \}$.)

(ii) The theory $\{p, q, r_1, r_2, r_3, p \rtimes q, q \rtimes p, q \rtimes r_1, r_1 \rtimes r_2, r_2 \rtimes r_3, r_3 \rtimes r_1, r_2 \rtimes r_2, r_3 \rtimes r_3\}$ is an example of a theory with a maximal dialectically preferred stage, for which no compatible maximal stage with larger or equal scope exists. The theory has one maximal stage and one maximal dialectically preferred stage, but they are not compatible. The theory's maximal stage M is specified by p, r_1 and the attack sentences of the theory: in M, p is justified, q defeated, r_1 justified, r_2 defeated and r_3 not taken into account. Its dialectically justified restriction is the dialectically preferred stage P specified by p and the attack sentences (in which r_1 and r_2 are not taken into account). The theory's maximal dialectically preferred stage MP, is specified by q and the attack sentences of the theory: in MP, p is defeated, q justified, r_1 defeated, and r_2 and r_3 not taken into account. It is its own dialectically justified restriction. M has larger scope than MP, but M's dialectically justified restriction has smaller scope than that of MP.

In the following figure, the inclusion and dialectically justified restriction maps between the stage types are summarized. The vertical arrows indicate the dialectically justified restriction maps, all of which are

[20] At a glance, this may seem to contradict a remark by Prakken & Vreeswijk (2002): '[...] it is easy to verify that preferred extensions correspond to maximal partial status assignments' (in their section on the approach by Bondarenko, Dung, Kowalski and Toni). However, although their notion of status assignment seems to be close to the notion of stage as it is used here (albeit that they use a restricted language, viz. only Dung sentences; cf. section 6 below), there is a crucial difference: in Prakken & Vreeswijk's status assignments *all* attacks available in the theory are taken into account. Instead in a DEFLOG stage it is possible that not all attacks are taken into account since in DEFLOG the attack information is treated in the same way as the other information.

surjective. The arrow from E to E indicates the identity map. All other arrows indicate inclusion maps. The ‘old’ stage types have been highlighted by the use of a bold font. MPCC, PCC, MDJ and CCDJ are the sets of stages that make the restriction maps surjective. There is no intuitive motivation behind the definition of these sets of stages. The only way in which these types of stages are relevant is that - surprisingly - *none of them* coincides with one of the previously defined types E, M, MP, P or CC. See Verheij (2000a) for further details.

$$\begin{tikzcd}
 & & & \underline{\mathbf{M}} \arrow[drr] \arrow[ddd, two heads] & & & \\
\textit{compatibility types} & \underline{\mathbf{E}} \arrow[urr, hook] \arrow[dr, hook] \arrow[ddd, Rightarrow] & & & & \underline{\mathbf{CC}} \arrow[r, hook] \arrow[ddd, two heads] & \underline{\mathbf{S}} \arrow[ddd, two heads] \\
 & & \underline{\mathrm{MPCC}} \arrow[rr, hook] \arrow[ddd, two heads] & & \underline{\mathrm{PCC}} \arrow[ur, hook] \arrow[ddd, two heads] & & \\
 & & & \underline{\mathrm{MDJ}} \arrow[drr] & & & \\
\textit{dialectical justification types} & \underline{\mathbf{E}} \arrow[urr, hook] \arrow[dr, hook] & & & & \underline{\mathrm{CCDJ}} \arrow[r, hook] & \underline{\mathbf{DJ}} \\
 & & \underline{\mathbf{MP}} \arrow[rr, hook] & & \underline{\mathbf{P}} \arrow[ur, hook] & &
\end{tikzcd}$$

The main lesson from the figure is that two ideas are very different: the idea that it is relevant to take as many sentences of a theory into account as possible (cf. the maximal stages) and the idea that it is relevant to defend against attack and incompatibility (cf. the dialectically preferred stages)

6 Dung’s argumentation frameworks and admissibility

Dung’s (1995) argumentation frameworks are a fruitful abstraction of ideas from nonmonotonic reasoning and logic programming. Here it is shown how Dung’s argumentation frameworks can be mimicked in DEFLOG. In fact, it is shown that Dung’s argumentation frameworks can be naturally regarded as DEFLOG theories that only use sentences of a subset of DEFLOG’s language. Since Dung has shown that his argumentation frameworks have close formal connections with well-established models of defeasible reasoning, such as Reiter’s (1980) default logic and logic programming, the results on DEFLOG presented here become of direct relevance for these models. Verheij (2000a) discusses relations of DEFLOG with previous research, such as with Reiter’s (1980) default logic, Vreeswijk’s (1993, 1997) abstract argumentation systems, Reason-Based Logic (Hage 1996, 1997, Verheij 1996b) and winning strategies in dialogue games (e.g., Prakken & Sartor 1997).

In this section it is also shown why Dung’s notion of admissibility cannot in general replace that of dialectical justification in the characterizations of the existence and multiplicity of dialectical interpretations (section 4).

Formally, an argumentation framework consists of a set, its elements called *arguments*, and a binary relation on that set, the *attack* relation. When (A, B) is in the attack relation, the argument A is said to attack B.

In Dung’s work, the notion of admissibility is central. It is closely related to DEFLOG’s dialectical justification. Using DEFLOG’s terminology, an argument C is *admissible* with respect to a theory Δ if C attacks any Δ-argument attacking it. Note that dialectical justification requires that all incompatible arguments are attacked, and not only the attacking arguments. The definition of admissibility given here depends of course on DEFLOG’s particular notions of argument and attack. There is however a straightforward way of mimicking Dung’s argumentation frameworks in DEFLOG for which this definition of admissibility is indeed an extrapolation of Dung’s admissibility, as follows.

Let each argument of an argumentation framework be an elementary sentence in DEFLOG’s language. Then an argumentation framework can be translated to a theory in DEFLOG by taking the union of the set of arguments in the framework and the set of sentences of the form $A \rightarrow \times B$, for any element (A, B) of the attack relation of the framework. Conversely, it is easy to restrict DEFLOG’s language in such a way that each theory in this restricted language corresponds to an argumentation framework in Dung’s sense: simply allow only elementary sentences and sentences of the form $\varphi \rightarrow \times\psi$, where φ and ψ are elementary. Let’s call sentences in this restricted sense *Dung sentences* and theories consisting of Dung sentences *Dung theories*.

It is now straightforward to check that several of Dung’s notions coincide with DefLog’s under this translation. Some care is needed however since certain terms have different meanings in Dung’s work and in DefLog. For instance, the use of the term ‘argument’ is different. Conflict-free sets of arguments (in Dung’s sense) correspond however with conflict-free sets of Dung sentences (in DefLog’s sense), Dung’s admissible sets of arguments correspond to the admissible arguments of Dung theories (in DefLog’s sense), and Dung’s stable extensions of argumentation frameworks correspond with DefLog’s dialectical interpretations of Dung theories. Verheij (2000a) formally establishes these results. The proofs are straightforward.

For theories using DefLog’s full language, dialectical justification and admissibility are easily seen to be different notions, but on the restricted language of Dung’s frameworks, the notions coincide:

Proposition (5.2.1)

Let Δ be a Dung theory. Then a Δ-argument is dialectically justifying with respect to Δ if and only if it is admissible with respect to Δ.

Proof: Dialectically justifying arguments are always admissible. (This does not depend on Δ being a Dung theory.) Let now C be an admissible argument, and let C’ be an argument incompatible with C. Since C and C’ consist of Dung sentences, the incompatibility of C and C’ implies that C attacks C’ or that C’ attacks C. In case C’ attacks C, also C attacks C’ since C is admissible. This shows that C is dialectically justifying.

Note that by this result the theorems on the existence and multiplicity of dialectical interpretations can *for Dung theories* be rephrased in terms of admissibility instead of dialectical justification. This is not the case for theories in general. Then the notion of dialectical justification is essential. The key point is that admissibility does not have all of the properties used in the proof of theorem (4.3) on the existence and multiplicity of dialectical interpretations. These properties are localization, union and separation at the base.

Their analogues for admissibility can be found by replacing ‘dialectically justifying’ by ‘admissible’ in the formulation of the properties. For instance, the union property (for pairs of arguments) for admissibility reads thus: if C and C’ are compatible arguments, that are admissible with respect to a theory Δ, then also $C \cup C’$ is admissible with respect to the theory. Separation at the base becomes (again for pairs of arguments): if C and C’ are incompatible arguments, that are admissible with respect to a theory Δ, then there are opposites φ and ψ in the theory, such that C supports φ and C’ supports ψ.

It is not hard to see that admissibility has the localization and union properties, but lacks the property of separation at the base.

For instance, that for admissibility, the property of separation at the base does not obtain, can be seen by inspection of the theory $\{p_1, p_1 \rightsquigarrow q, p_2, p_2 \rightsquigarrow (q \rightsquigarrow \times q)\}$. With respect to the theory, there are four admissible arguments with a maximal number of elements, viz. each three-element subset of the theory. (Note that each argument of the theory is admissible since there are no attacking arguments.) Any pair of these arguments is incompatible, yet there is no sentence that is defeated by an argument, let alone by an admissible argument, as is required by the property of separation at the base.

It follows straightforwardly that the localization property obtains for admissibility: since, when E is a dialectical interpretation of a theory Δ, $J(E) \cap \Delta$ is dialectically justifying with respect to Δ, $J(E) \cap \Delta$ is certainly admissible.

The proof of the union property for admissibility is almost trivial since any attack of the union of a collection of arguments is also an attack of one of the arguments in the collection.

Inspection of the proof of theorem (4.3) shows that the property of separation at the base is only used in the ‘if’-part. The ‘only if’-part indeed has an analogue for admissibility since it only uses localization and union. The theory $\{p_1, p_1 \rightsquigarrow q, p_2, p_2 \rightsquigarrow (q \rightsquigarrow \times q)\}$ (the counterexample against the property of separation at the base) shows that the analogue of the ‘if’-part is in fact not true. All sentences in the theory are ‘admissibly justifiable’, i.e., supported by an admissible argument, since any argument of the theory is admissible. No sentence in the theory is ‘admissibly defeasible’, i.e., attacked by an admissible argument, since there is no attacking argument at all. Still, the theory has no dialectical interpretation.

Verheij (2000a) expands this meta-analysis for other results (e.g., concerning so-called dialectically preferred and admissibly preferred arguments, i.e., those dialectically justifying or admissible arguments that are maximal with respect to set inclusion) and for other notions that are similar to dialectical justification.

Bondarenko, Dung, Kowalski & Toni (1997) have used admissibility in their discussion of an abstract, argumentation-theoretic approach to default reasoning. Their setting is just as Dung’s (1995) related to

DEFLOG's, yet they focus on deductive systems. Interestingly, whereas in DEFLOG dialectical negation × is treated as an ordinary connective, Bondarenko, Dung, Kowalski & Toni consider the question which sentences are the contraries of others as part of the domain theory (as the mapping from sentences to their contraries is explicitly represented in their assumption-based frameworks). It seems that the notion of dialectical justification can be directly transplanted to their system. For the reasons, discussed here and in section 4, it can be expected that dialectical justification has better properties for analyzing assumption-based frameworks than admissibility.

7 Conclusion

In this paper, a logic has been presented that shows how sets of prima facie justified assumptions can be interpreted. For this purpose, the notion of dialectical interpretation has been introduced.

When theories are interpreted dialectically, some prima facie justified assumptions are actually justified and others defeated. More theories are interpretable dialectically than 'monolectically', i.e., as sets of sentences assumed to be all true. In other words, there are more theories with dialectical interpretations than theories with models.

A fundamental complication of dialectical interpretation of theories in terms of dialectical interpretations is that theories can have zero, one or several dialectical interpretations. This complication is common for nonmonotonic logics. The existence problem asks for a necessary and sufficient criterion for the existence of a dialectical interpretation of a theory. The multiplicity problem asks for a necessary and sufficient criterion for the existence of multiple dialectical interpretations of a theory. The notion of dialectical justification, introduced in the present paper, gives rise to necessary and sufficient criteria that solve the existence and multiplicity problems for dialectical interpretations. An argument is dialectically justifying when it attacks all arguments that are incompatible with it. The properties of dialectical justification, especially the union, localization and separation properties, make it particularly suitable for the analysis of dialectical interpretations. The idea is that a dialectical interpretation exists if and only if there is a part of the theory in the context of which no sentence of the theory is dialectically ambiguous (i.e., both dialectically justifiable and dialectically defeasible), while all sentences of the theory are dialectically interpretable (i.e., either dialectically justifiable or dialectically defeasible) in the context of that part of the theory. Multiple dialectical interpretations exist if and only if there are multiple incompatible parts with these properties.

DEFLOG uses a simple, dialectically interpreted logical language using ordinary connectives × and $\leadsto$ that is suitable for the analysis of central topics of dialectical argumentation, such as Toulmin's argument scheme, Pollock's rebutting and undercutting defeaters. An important consequence of the choice of language is that in DEFLOG all information concerning justification and defeat is expressible in the logical object language as contingent information. There is no need for separate classes of defeasible rules of inference, priority information or pre-defined conclusive force relations between arguments. All these kinds of information can be expressed directly in DEFLOG's language, along with the other contingent information.

The idea of stages provides a different approach towards the investigation of the local properties of dialectical interpretation. A theory's stages are the dialectical interpretations of parts of the theory. Instead of maximizing only the justified assumptions of a theory in a stage, it is also possible to maximize the whole set of interpreted assumptions of a theory. It turns out that the types of maximization are perpendicular, in the sense that maximization in one sense does not imply maximality in the other sense. The result is a plethora of types of stages, with few interrelations.

This suggests that one should not consider each as a different type of semantics, as is for some types suggested in the work of Dung (1995) and Bondarenko, Dung, Kowalski & Toni (1997) and also in Prakken & Vreeswijk's overview (2002), but merely as partial interpretations with an interesting special property. In other words, to me, there is only one 'genuine' dialectical semantics, viz. dialectical interpretation, a variant of stable semantics. All other notions, such as compatibility classes, dialectically preferred stages and maximal stages, are in the first place tools in the investigation of the properties of dialectical interpretations. The use of the notion of dialectical justification in the existence and multiplicity problems is an example of the application of such tools.

Acknowledgments

The author gladly acknowledges the financial support by the Dutch national Programme Information Technology and Law (ITeR) for the research reported in this paper (project number 01437112). The author thanks Phan Minh Dung, Jaap Hage, David Hitchcock, Robert Kowalski, Peter McBurney, Henry Prakken, Bram Roth and Gerard Vreeswijk for comments and discussion.

References

Bench-Capon, T. (1995). Argument in Artificial Intelligence and Law. *Legal knowledge based systems. Telecommunication and AI & Law* (eds. J.C. Hage, T.J.M. Bench-Capon, M.J. Cohen & H.J. van den Herik), pp. 5-14. Koninklijke Vermande, Lelystad.

Bench-Capon, T.J.M. (1997). Argument in Artificial Intelligence and Law. *Artificial Intelligence and Law*, Vol. 5, pp. 249-261.

Bondarenko, A., Dung, P.M., Kowalski, R.A., & Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, Vol. 93, pp. 63-101.

Chesñevar, C.I., A.G. Maguitman & R.P. Loui (2000). Logical models of argument. *ACM Computing Surveys*, Vol. 32, No. 4, pp. 337-383.

Dung, P.M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, Vol. 77, pp. 321-357.

Eemeren, F.H. van, Grootendorst, R., & Snoeck Henkemans, F. (1996). *Fundamentals of Argumentation Theory. A Handbook of Historical Backgrounds and Contemporary Developments.* Lawrence Erlbaum Associates, Mahwah (New Jersey).

Gabbay, D.M., Hogger, C.J., & Robinson, J.A. (eds.) (1994). *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Nonmonotonic Reasoning and Uncertain Reasoning*. Clarendon Press, Oxford.

Gelfond, M. & Lifschitz, V. (1988). The stable model semantics for logic programming. *Logic Programming. Proceedings of the Fifth International Conference and Symposium* (eds. R.A. Kowalski & K.A. Bowen), pp. 1070-1080. The MIT Press, Cambridge (Massachusetts).

Ginsberg, M.L. (ed.) (1987). *Readings in Nonmonotonic Reasoning.* Morgan Kaufmann Publishers, Los Altos (California).

Haack, S. (1978). *Philosophy of logics.* Cambridge University Press, Cambridge.

Hage, J.C. (1997). *Reasoning with Rules. An Essay on Legal Reasoning and Its Underlying Logic.* Kluwer Academic Publishers, Dordrecht.

Lin, F. (1993). An argument-based approach to nonmonotonic reasoning. *Computational Intelligence*, Vol. 9, No. 3, pp. 254-267.

Loui, R.P. (1998). Process and Policy: Resource-Bounded Non-Demonstrative Reasoning. *Computational Intelligence*, Vol. 14, No. 1.

Nute, D. (1994). Defeasible Logic. *Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Nonmonotonic Reasoning and Uncertain Reasoning* (eds. D.M. Gabbay, C.J. Hogger & J.A. Robinson), pp. 353-395. Clarendon Press, Oxford.

Pollock, J.L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person.* The MIT Press, Cambridge (Massachusetts).

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, Vol. 36, pp. 27-47.

Prakken, H. (1997). *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law.* Kluwer Academic Publishers, Dordrecht.

Prakken, H., & Sartor, G. (1996). A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning. *Artificial Intelligence and Law*, Vol. 4, pp. 331-368.

Prakken, H. & Vreeswijk, G.A.W. (2002). Logics for Defeasible Argumentation. *Handbook of Philosophical Logic, Second Edition* (eds. D.M. Gabbay & F. Guenthner), Vol. 4, pp. 218-319. Kluwer Academic Publishers, Dordrecht.

Read, S. (1995). *Thinking About Logic. An Introduction to the Philosophy of Logic.* Oxford University Press, Oxford.

Reiter, R. (1980). A Logic for Default Reasoning. *Artificial Intelligence*, Vol. 13, pp. 81-132.

Rescher, N. (1964). *Hypothetical reasoning*. North-Holland Publishing Company, Amsterdam.

Toulmin, S.E. (1958). *The uses of argument.* University Press, Cambridge.

Verheij, B. (1996a). Two approaches to dialectical argumentation: admissible sets and argumentation stages. *NAIC'96. Proceedings of the Eighth Dutch Conference on Artificial Intelligence* (eds. J.-J.Ch. Meyer & L.C. van der Gaag), pp. 357-368. Also presented at the *Computational Dialectics Workshop* at FAPR-96. June 3-7, 1996, Bonn.

Verheij, B. (1996b). *Rules, Reasons, Arguments. Formal studies of argumentation and defeat*. Dissertation. Universiteit Maastricht, Maastricht.

Verheij, B. (1999). Automated Argument Assistance for Lawyers. *The Seventh International Conference on Artificial Intelligence and Law. Proceedings of the Conference*, pp. 43-52. ACM, New York (New York).

Verheij, B. (2000a). DEFLOG - a logic of dialectical justification and defeat. Manuscript available at http://www.metajur.unimaas.nl/~bart/publications.htm.

Verheij, B. (2000b). Dialectical Argumentation as a Heuristic for Courtroom Decision-Making. *Rationality, Information and Progress in Law and Psychology. Liber Amicorum Hans F. Crombag* (eds. Peter J. van Koppen & Nikolas H.M. Roos), pp. 203-226. Metajuridica Publications, Maastricht.

Verheij, B. (2001a). Evaluating arguments based on Toulmin's scheme. *OSSA 2001: Argumentation and its applications (The Ontario Society for the Study of Argumentation).*

Verheij, B. (2001b). Legal decision making as dialectical theory construction with argumentation schemes. *The 8th International Conference on Artificial Intelligence and Law. Proceedings of the Conference*, pp. 225-226. ACM, New York.

Verheij, B. (2002). On the existence and the multiplicity of extensions in dialectical argumentation. *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR'2002)* (eds. S. Benferhat & E. Giunchiglia), pp. 416-425. Toulouse.

Verheij, B. (*to appear*). Artificial argument assistants for defeasible argumentation. *Artificial Intelligence*.

Vreeswijk, G. (1997). Abstract argumentation systems. *Artificial Intelligence*, Vol. 90, pp. 225-279.